SUPPORTING UNDERGRADUATE RESEARCH: RECOMMENDING PERSONALIZED RESEARCH PROJECTS TO UNDERGRADUATES

Yang Liu, School of Management, University of Science and Technology of China, Hefei, Anhui, PR China; Department of Information Systems, City University of Hong Kong, Kowloon, Hong Kong, PR China, liuy7@mail.ustc.edu.cn

Jian Ma, Department of Information Systems, City University of Hong Kong, Kowloon, Hong Kong, PR China, isjian@cityu.edu.hk

Wei Du, Department of Information Systems, City University of Hong Kong, Kowloon, Hong Kong, PR China, weidu7-c@my.cityu.edu.hk

Chen Yang, Department of Management Science, Shenzhen University , Shenzhen, Guangdong, PR China, yangc0201@gmail.com

Zhongsheng Hua, School of Management, University of Science and Technology of China, Hefei, Anhui, PR China, zshua@ustc.edu.cn

Abstract

Undergraduates' participation in faculty-mentored research is becoming an important issue in tertiary education in recent years, and it benefits both undergraduates and faculty members. In reality, many faculty members have research projects that need help from undergraduates, but undergraduates can hardly find the information, which creates information asymmetry problem. Besides, undergraduates lack the experience of doing academic research, the research interest information is incomplete, so they have difficulties in choosing suitable research projects. Thus recommender systems are necessary to facilitate undergraduates' participation in research projects. Traditional recommendation approaches require relative complete information for decision making, and they can hardly meet the requirements as undergraduates' research information is incomplete. In this study, we propose a two-stage model that integrates content-based method with collaborative method by leveraging research social networks, where undergraduates are encouraged to connect with faculty members and participate in social network activities, through which research information is collected. The proposed two-stage model alleviates the problems of information asymmetry and incomplete information. The recommender system has been developed in ScholarMate (www.scholarmate.com), and it allows undergraduates to choose suggested research projects.

Keywords: undergraduate research, research project recommendation, content-based methods, research social networks.

1 INTRODUCTION

Undergraduate research plays an important role in academic research and innovation. For most research universities in China, in order to nurture the research culture within undergraduates and cultivate young power in academic community, undergraduates are provided with opportunities to work closely with faculty members (some are even world-class scholars) and participate in academic activities. Many research universities in developed countries (e.g., the United States), have encouraged undergraduates to participate in faculty-mentored research since 1969 (Merkel 2001). Such collaboration benefits both undergraduates and faculty members (Hunter et al. 2007). The participation in research projects offers undergraduates the chance to learn new skills, gain confidence, and prepare for future careers (Elgren & Hensel 2006; Russell et al. 2007).Through collaboration, faculty members obtain extra help and have better chances to try out new ideas or speculative experiments (Goodlad 1998).

In information era, people are facing larger amounts of information than ever before. At the same time, when making a choice, they usually lack decisive first-hand information to support their decision. For undergraduates, they usually do not have the experience of doing academic research, and their research interest information is incomplete. When choosing the research project, undergraduates tend to ask help from other senior undergraduates who have participated in the projects before or search online. These ways would be restricted to the friends of undergraduates and the limited information they have searched. There exist incomplete information, information asymmetry problem and the mismatch issue between undergraduates and projects. Overall, undergraduates have difficulties in finding proper research projects. There is a pressing need to develop an effective and efficient approach that helps undergraduates find proper research projects.

Several methods have been proposed to solve the problems in undergraduate research. Integrated web database allows academic members to post research project information and allows undergraduates to search for suitable research projects (Snow et al. 2010). The usage of web database has transformed the process of research projects selecting from offline connections to online search, it addresses the information asymmetry problem partially. However, it is only an online platform that helps undergraduates get information, and undergraduates make the decision by themselves after collecting the information, thus undergraduates still have difficulties in finding proper research projects. Recommender systems have been proposed in the field of education (Manouselis et al. 2011). Recommender systems could help undergraduates find the research projects they would like, but the personal demands of undergraduates are neglected. The aim of this research is to help undergraduates find research projects they are really interested in with minimal search efforts.

For recommender systems, rating-based approach is the most popular approach (Adomavicius & Tuzhilin 2005), where the recommendation is based on the explicit or implicit ratings of users. Rating-based approaches can be classified into content-based, collaborative and hybrid methods (Balabanović & Shoham 1997). Content-based methods in research projects recommendation are based on the profiles of undergraduates and projects, they are appropriate to recommend text-based items (i.e., projects). When lacking decisive information to make a choice, it is a good strategy to choose as other like-minded, similarly-situated people have successfully chosen in the past (Hill et al. 1995). However, collaborative methods have data sparsity problem, they need a critical mass of users (Adomavicius & Tuzhilin 2005). Hybrid methods combine the two methods. Overall, Traditional methods require relative complete information and overlook the importance of research social network features.

Hence in this paper we propose a two-stage recommendation model to recommend highly relevant and socially endorsed research projects to undergraduates. It is a hybrid recommendation approach, combining content-based methods and collaborative methods. At the first stage, by finding relevant research projects, it removes irrelevant projects. With the rapid development of research social networks, people tend to find research information on research social networks (Noorden 2014), we also use research social network approach and web usage mining to increase the number of ratings. At the second stage, by finding the projects that the friends with similar background have participated in,

the recommendation results are more reliable. The usage of activities in research social networks and connections can mine more information. Thus, the incomplete information and information asymmetry problem can be alleviated by the proposed approach.

The rest of paper is organized as follows. Related work in information recommendation services can be found in Section 2. The third section describes the two-stage undergraduate research projects recommendation method. The fourth section shows the interfaces of the implemented system and evaluation design. The last section is the summary of our research and the future work.

2 RELATED WORK

Undergraduate research projects recommendation belongs to information recommendation services. So we review the recommendation approaches in similar scenarios, and then summarize the deficiencies of current approaches.

Information recommendation services have been applied in various situations, such as learning materials recommendation (Ghauth & Abdullah 2010; Salehi 2013), recommending R&D projects to reviewers (Silva et al. 2013) and job recommendation (Malinowski et al. 2006; Paparrizos et al. 2011). In e-learning environments, Ghauth and Abdullah (2010) investigated learning materials recommendation problem by incorporating good learners' ratings into recommender systems. They chose content-based methods as recommendation approaches. A system that combining implicit and explicit attributes based collaborative filtering and BI-Directional Extension based frequent closed sequence mining was developed (Salehi 2013). In the R&D projects-reviewers recommendation situation, Silva et al. (2013) proposed a social network-empowered research analytics framework (named RAF) to recommend R&D projects. RAF considered the matching of research relevance, social connections and productivity. In the match between jobs and persons, a bilateral recommendation approach considering the preferences of the recruiters and the candidates was proposed (Malinowski et al. 2006). Paparrizos et al. (2011) addressed the problem of recommending suitable jobs to people by machine learning methods, and they used past job profiles and institution profiles of the users.

In the situations of learning materials-users, R&D projects-reviewers and jobs-persons recommendations, users, reviewers and persons have the experience of reading learning materials, reviewing R&D projects and having jobs, thus the information is relative complete. However, undergraduates do not have the experience of doing academic research, for them, the information is relative incomplete. Existing methods are insufficient for personalized research projects recommendation.

Based on the related literatures, there are several research gaps. First, existing approaches require relative complete information. While for undergraduates, the research interest information is incomplete. Existing approaches are not suited in this situation. Second, existing studies overlook the activities in research social networks. With the rapid development of research social networks, researchers and students tend to find research information, post content, and discuss in research social networks (Noorden 2014). By tracking the activities in research social networks, the potential preference of undergraduates can be obtained. Thus, we propose a two-stage model that integrates content-based method with collaborative method by leveraging research social networks.

3 PROPOSED APPROACH

3.1 Overview of the Proposed Approach

In this study, we propose a two-stage model by leveraging research social networks to recommend proper research projects to undergraduates. Figure 1 indicates the mechanism of the approach.

In the proposed approach, there are two stages: filtering stage and ranking stage. At the filtering stage, we use content-based method to select relevant research projects. Besides the explicit ratings that can be extracted from undergraduate database and project web database, we utilize web usage mining to

extract implicit ratings from undergraduates' activities in ScholarMate (i.e., an online research social network). At the ranking stage, if the relevant research project is the project that friends of undergraduates have successfully participated in, the project will be highly recommended to the undergraduates. Finally, the two-dimensional scores are aggregated. Based on the aggregated results, the final list can be obtained.

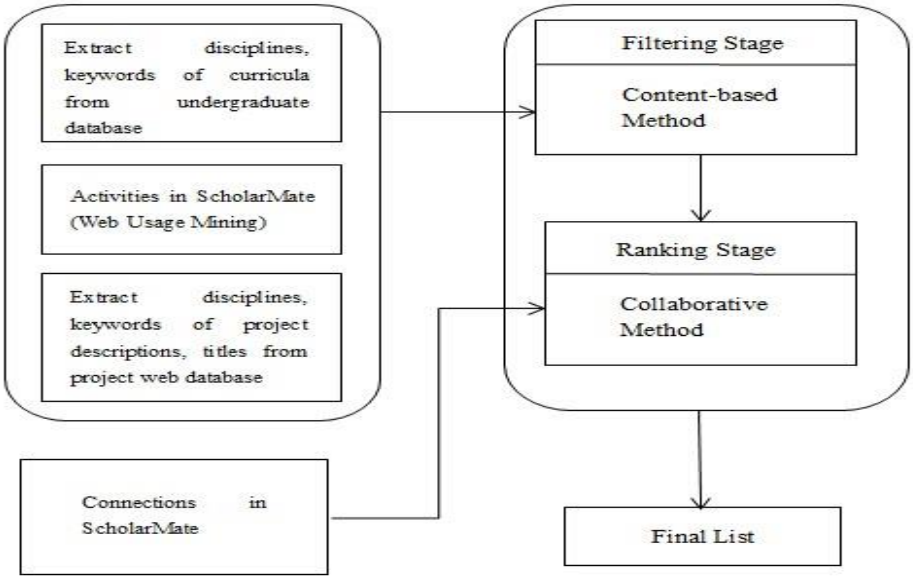

Figure 1. The framework of the proposed approach

3.2 Content-based Filtering

3.2.1 Explicit and Implicit Ratings with Web Usage Mining

Profiling is the process of determining key attributes of users and items. In the research projects-undergraduates recommendation situation, we need to extract key features of undergraduates and research projects first.

For selecting research projects, the knowledge background of undergraduates is important. They usually prefer the projects which are similar to their current major or the projects they are interested in. So we extract disciplines, keywords of curricula that undergraduates have taken from undergraduate database. Disciplines, keywords of project descriptions and titles can be extracted from project web database to describe research projects.

There are activities in research social networks, which indicate the potential preference of undergraduates. The activities in ScholarMate include liking, sharing, commenting, downloading papers and clicking. The preference of undergraduates is measured by counting the number of occurrence of URLs mapped to the research area from the clickstream of the users (Cho & Kim 2004). Let p_{ij}^{l} be the total number of occurrences of liking of an undergraduate *i* in research area *j*. Likewise, p_{ij}^{s}, p_{ij}^{co}, p_{ij}^{d} and p_{ij}^{cl} are defined as the total number of occurrences of sharing, commenting, downloading and clicking of an undergraduate *i* in research area *j*. We obtain the preference of undergraduates by Equation 1.

$$p_{ij} = \frac{p_{ij}^{l} - \min_{1\le j\le|A|}\left(p_{ij}^{l}\right)}{\max_{1\le j\le|A|}\left(p_{ij}^{l}\right) - \min_{1\le j\le|A|}\left(p_{ij}^{l}\right)} + \frac{p_{ij}^{s} - \min_{1\le j\le|A|}\left(p_{ij}^{s}\right)}{\max_{1\le j\le|A|}\left(p_{ij}^{s}\right) - \min_{1\le j\le|A|}\left(p_{ij}^{s}\right)} + \frac{p_{ij}^{co} - \min_{1\le j\le|A|}\left(p_{ij}^{co}\right)}{\max_{1\le j\le|A|}\left(p_{ij}^{co}\right) - \min_{1\le j\le|A|}\left(p_{ij}^{co}\right)} + \frac{p_{ij}^{d} - \min_{1\le j\le|A|}\left(p_{ij}^{d}\right)}{\max_{1\le j\le|A|}\left(p_{ij}^{d}\right) - \min_{1\le j\le|A|}\left(p_{ij}^{d}\right)} + \frac{p_{ij}^{cl} - \min_{1\le j\le|A|}\left(p_{ij}^{cl}\right)}{\max_{1\le j\le|A|}\left(p_{ij}^{cl}\right) - \min_{1\le j\le|A|}\left(p_{ij}^{cl}\right)} \quad (1)$$

Where the total number of undergraduates is *U*, the total number of research area is *A*. p_{ij} is a sum of normalized value of p_{ij}^{l}, p_{ij}^{s}, p_{ij}^{co}, p_{ij}^{d} and p_{ij}^{cl}. It ranges from 0 to 5, and the larger the value, the more preferred the research area is. When the user is interested in some papers, they may not only download them, but also click, like, comment on them, or share them to his/her social networks. So the Equation 1 reflects the different weights.

From the activities in ScholarMate, we can know the research areas that undergraduates are interested in, and we extract keywords from the interested research area.

3.2.2 Semantic Profile Matching

Keywords which are extracted from project descriptions and titles can be described as $P_p = <(K_{p1,}W_{p1}),(K_{p2,}W_{p2})...(K_{pm,}W_{pm})>$. P_p denotes the keywords set of research project p, $K_{p1},K_{p2}\ldots K_{pm}$ are the keywords of project p, $W_{p1},W_{p2},\ldots W_{pm}$ are the weights of them respectively, and they indicate the times of occurrence. Keywords which represent the preference of undergraduates can be extracted from the curricula and the interested research area. The keywords and weights can be described as $C_c = <(k_{c1,}w_{c1}),(k_{c2,}w_{c2})...(k_{cm,}w_{cm})>$. C_c denotes the keywords set of curriculum c and interested research area, $k_{c1},k_{c2}\ldots k_{cm}$ are the keywords of curriculum c and interested research area, $w_{c1},w_{c2},\ldots w_{cm}$ are the weights of them respectively, and they indicate the times of occurrence. To make it convenient to compute, we set the number of pm and cm as the same value.

The larger the same part of keywords between research project and curriculum, interested research area, the more relevant the research project is. For keywords $K_{p1},K_{p2}\ldots K_{pm}$, the weights from research projects are $W_{p1},W_{p2}\ldots W_{pm}$, the weight from curricula and interested research area are $w_{c1},w_{c2}\ldots w_{cm}$.So the relevance of research project is denoted by the score in Equation 2.

$$Score(key_m) = \begin{cases} 0, \ no \quad keywords \quad matched \\ 1, \ exist \quad matched \quad keywords, \quad w_{cm} >= W_{pm} \\ w_{cm}/W_{pm}, \ exist \quad matched \quad keywords, \quad w_{cm} < W_{pm} \end{cases} \quad (2)$$

The larger the score, the more relevant the research project is.

3.2.3 The Relatedness of the Disciplines

In direct keyword matching methods, there are problems of polysemy and synonyms (Pazzani & Billsus 2007). In order to increase the accuracy, we also measure the relatedness of the disciplines. The disciplines of undergraduates can be extracted from undergraduate database and the research area they are interested in from ScholarMate. The disciplines of research projects can be extracted from the projects database.

The discipline category is based on the Chinese Ministry of Education (MOE). According to the category tree (Li & Shiu 2012) and WordNet similarity (Pedersen et al. 2004), the similarity between two different disciplines can be obtained by Equation 3.

$$Score\left(d_k^u, d_h^p\right) = 2 \cdot d\left(LCS\left(d_k^u, d_h^p\right)\right) / \left(d\left(d_k^u\right) + d\left(d_h^p\right)\right) \quad (3)$$

In the Equation 3, the discipline of undergraduate *k* is denoted as d_k^u, the discipline of research project *h* is denoted as d_h^p. $d\left(d_k^u\right)$ means the length from the root to d_k^u. $LCS\left(d_k^u, d_h^p\right)$ denotes the least common subsume between d_k^u and d_h^p.

The total matching score of relevance can be obtained in Equation 4.

$$Score\left(Rel\right) = Score\left(key_m\right) + Score\left(d_k^u, d_h^p\right) \quad (4)$$

Therefore, the relevant research projects can be produced.

3.3 Collaborative Method

If the friends of undergraduates have successfully participated in the research projects, the undergraduate will also prefer them. So among the relevant research projects, if the projects are the ones that friends have joined, the projects will be highly recommended. The score of connectivity is indicated in Equation 5.

$$Score(Con) = N_j / \max N_j \quad (5)$$

$Score(Con)$ denotes the score of connectivity among undergraduates. N_j is the number of friends who have joined the undergraduate research projects.

The information of friends with similar background can be obtained from the research social network ScholarMate.

3.4 Aggregation and Ranking

In the process of matching, relevance matching plays the role of filtering, while connectivity matching plays the role of awarded marks. The aggregation can be described as:

$$Score = \mu Score(\mathrm{Re}l) + \lambda Score(Con) \quad (6)$$

In Equation 6, $\mu > \lambda$. This is because the weight of relevance is larger than the weight of connectivity. At last, the final recommendation list can be produced; undergraduates can make their own decision with considering the result of the recommendation.

4 SYSTEM IMPLEMENTATION

ScholarMate is an online research social network, most of the users are students and researchers, and they communicate and cooperate with each other. In ScholarMate, most of the users are Chinese, they share their papers and projects. Others can comment on them, share them to their social networks, express their like and download. ScholarMate provides the platform and channel for researchers to contact with each other without the limitation of space. There is research CV function in ScholarMate. The figure 2 shows the interface of CV and recommendation. Based on the basic information of undergraduates, research projects, social network activities and connections, undergraduate users in ScholarMate can get recommendation of research projects. It helps undergraduates make decisions and find suitable research projects.

To verify the effectiveness of the approach, we will make a recommendation result comparison between the proposed approach and baseline methods. The evaluation metrics include the average rating score and normalized discounted cumulative gain.

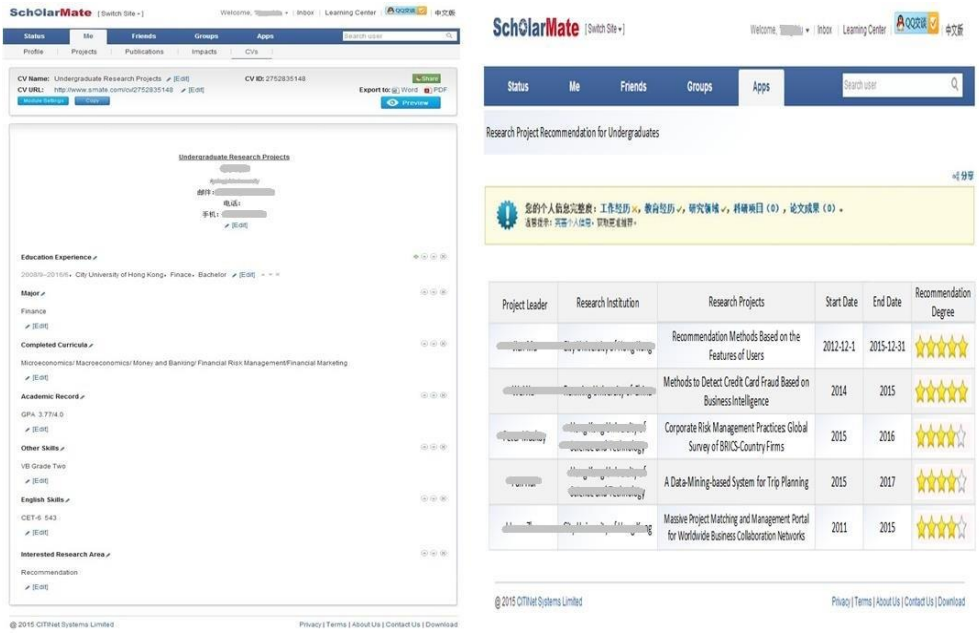

Figure 2. The interface of CV and recommendation system

5 CONCLUSIONS AND FUTURE WORK

In this paper, the two-stage model has been proposed for undergraduate research. In order to mine the preference of undergraduates, social network approach and web usage mining are proposed to increase the rating number and generate more profiles, based on which we further apply content-based method and collaborative method to recommend highly relevant and socially endorsed research projects. The recommender system has been developed and implemented in ScholarMate, an online research social network platform. The proposed system encourages undergraduates to connect with researchers, peers, and participate in network research activities. It collects undergraduate research social network activity information and other basic information, and then recommends suitable research projects to undergraduates.

There are two main contributions of this paper. First, we propose a two-stage model that integrates content-based method with collaborative method by leveraging research social networks. Particularly, in the content-based method part, we use implicit and explicit ratings. Undergraduates are encouraged to connect with researchers, peers, and participate in social network activities, through which research information is collected, and potential preference of undergraduates is mined. It enriches the recommendation literatures in which user's information is incomplete. Second, a recommender system is designed and implemented to facilitate undergraduate research, and it allows undergraduates to choose suggested projects. Experiments will be conducted to evaluate the effectiveness of the proposed method and the recommender system in the future.

Acknowledgements

This work was supported by the General Research Fund of Hong Kong Research Grant Council (CityU 148012) and National Natural Science Foundation of China (71371164).

References

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. Knowledge and Data Engineering, IEEE Transactions on, 17(6), 734-749.

Ash Merkel, C. (2003). Undergraduate research at the research universities.New Directions for Teaching and Learning, 2003(93), 39-54.

Balabanović, M. & Shoham, Y. (1997). Fab: content-based, collaborative recommendation. Communications of the ACM 40(3): 66-72.

Cho, Y. H., & Kim, J. K. (2004). Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce. Expert systems with Applications, 26(2), 233-246.

Elgren, T., & Hensel, N. (2006). Undergraduate research experiences: Synergies between scholarship and teaching. Peer Review, 8(1), 4.

Ghauth, K. I., & Abdullah, N. A. (2010). Learning materials recommendation using good learners' ratings and content-based filtering. Educational Technology Research and Development, 58(6), 711-727.

Goodlad, S. (1998). Research opportunities for undergraduates. Studies in Higher Education, 23(3), 349-356.

Hill, W., Stead, L., Rosenstein, M., & Furnas, G. (1995, May). Recommending and evaluating choices in a virtual community of use. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 194-201). ACM Press/Addison-Wesley Publishing Co.

Hunter, A. B., Laursen, S. L., & Seymour, E. (2007). Becoming a scientist: The role of undergraduate research in students' cognitive, personal, and professional development. Science Education, 91(1), 36-74.

Li, Y.M., & Shiu, Y.L. (2012). A diffusion mechanism for social advertising over microblogs. Decision Support Systems 54(1): 9-22.

Malinowski, J., Keim, T., Wendt, O., & Weitzel, T. (2006, January). Matching people and jobs: A bilateral recommendation approach. In System Sciences, *2006.* HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on (Vol. 6, pp. 137c-137c). IEEE.

Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H., & Koper, R. (2011). Recommender systems in technology enhanced learning. *Recommender systems handbook* (pp. 387-415). Springer US.

Merkel, C. A. (2001). Undergraduate research at six research universities.Pasadena, CA: California Institute of Technology.

Paparrizos, I., Cambazoglu, B. B., & Gionis, A. (2011, October). Machine learned job recommendation. In Proceedings of the fifth ACM Conference on Recommender Systems (pp. 325-328). ACM.

Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In The adaptive web (pp. 325-341). Springer Berlin Heidelberg.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004, May). WordNet:: Similarity: measuring the relatedness of concepts. In Demonstration papers at HLT-NAACL 2004 (pp. 38-41). Association for Computational Linguistics.

Russell, S. H., Hancock, M. P., & McCullough, J. (2007). Benefits of undergraduate research experiences. Science(Washington), 316(5824), 548-549.

Salehi, M. (2013). Application of implicit and explicit attribute based collaborative filtering and BIDE for learning resource recommendation. Data & Knowledge Engineering, 87, 130-145.

Silva, T., Guo, Z., Ma, J., Jiang, H., & Chen, H. (2013). A social network-empowered research analytics framework for project selection. Decision Support Systems, 55(4), 957-968.

Snow, A. A., de Cosmo, J., & Shokair, S. M. (2010). Low-Cost Strategies for Promoting Undergraduate Research at Research Universities. Peer Review, 12(2), 16-19.

Van Noorden, R. (2014). "Online collaboration: Scientists and the social network." Nature,512(7513): 126-129.